

Systematic analysis of the gerontome reveals links between aging and age-related diseases

Maria Fernandes^{1,2}, Cen Wan^{3,4}, Robi Tacutu¹, Diogo Barardo¹, Ashish Rajput⁵, Jingwei Wang¹, Harikrishnan Thoppil¹, Daniel Thornton¹, Chenhao Yang¹, Alex Freitas³, João Pedro de Magalhães^{1,*}

¹Integrative Genomics of Ageing Group, Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool, UK.

²LaSIGE - Large-Scale Informatics Systems Laboratory, Faculty of Sciences, University of Lisbon, Portugal.

³School of Computing, University of Kent, Canterbury, United Kingdom.

⁴Department of Computer Science, University College London, London, United Kingdom.

⁵Research Group for Computational Systems Biology, German Center for Neurodegenerative Diseases (DZNE), Göttingen, Germany.

* Corresponding author: Telephone: +44 151 7954517; Fax: +44 151 795 8420; E-mail: jp@senescence.info

© The Author 2016. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License

(<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

In model organisms, over 2,000 genes have been shown to modulate aging, the collection of which we call the “gerontome”. Although some individual aging-related genes have been the subject of intense scrutiny, their analysis as a whole has been limited. In particular, the genetic interaction of aging and age-related pathologies remains a subject of debate. In this work, we perform a systematic analysis of the gerontome across species, including human aging-related genes. First, by classifying aging-related genes as pro- or anti-longevity, we define distinct pathways and genes that modulate aging in different ways. Our subsequent comparison of aging-related genes with age-related disease genes reveals species-specific effects with strong overlaps between aging and age-related diseases in mice, yet surprisingly few overlaps in lower model organisms. We discover that genetic links between aging and age-related diseases are due to a small fraction of aging-related genes which also tend to have a high network connectivity. Other insights from our systematic analysis include assessing how using datasets with genes more or less studied than average may result in biases, showing that age-related disease genes have faster molecular evolution rates and predicting new aging-related drugs based on drug-gene interaction data. Overall, this is the largest systems-level analysis of the genetics of aging to date and the first to discriminate anti- and pro-longevity genes, revealing new insights on aging-related genes as a whole and their interactions with age-related diseases.

Introduction

Aging is a major social and medical challenge of the 21st century. The most accepted mechanisms of aging include inflammation (1), apoptosis, oxidative stress, accumulation of DNA damage, cell cycle deregulation, and mitochondrial dysfunction (2-4). In addition, one of the major breakthroughs in the field of aging research is the discovery that, in model organisms, aging is under genetic regulation (5). In the past 20 years, aging has been shown to be under genetic control in various short-lived model organisms, and in particular in yeast, worms, flies and mice. According to the GenAge database (6), over 2,000 genes can modulate aging and/or longevity in model organisms. We call the collection of these aging-related genes the “gerontome” (7). Many of these genes work in common pathways (4), which include the insulin-like growth factor (IGF-1) signalling pathway, the target of rapamycin (TOR) pathway and the AMP kinase pathway (5).

Although some individual aging-related genes have been the subject of intense scrutiny, their analysis as a whole has been limited (8-11). Yet genes and proteins do not act individually. Therefore, biological networks provide a more realistic description of biological systems than single-molecule studies and give way to the integration of several types of data (12). Indeed, network analyses have already revealed insights on aging and its manipulation (13-15).

Aging is associated with various diseases. The main categories of aging-related pathologies are: cancer, cardiovascular diseases, neurodegenerative diseases, nutritional and metabolic diseases (16-18). The relationship between aging and age-related diseases has long been a contentious topic. A previous study has shown that the analysis of networks can uncover links between aging-related genes and age-related diseases (19), but many questions remain unanswered, like which aging-related genes and pathways are important in these interactions?

Moreover, we have further classified aging-related genes as anti- or pro-longevity, depending on how they are genetically manipulated and whether they increase or decrease lifespan in model organisms (6). Whether and how anti- and pro-longevity genes interact with aging disease-related genes is unknown.

In this work, we performed a systematic analysis of the gerontome, the largest such analysis to date and the first to discriminate anti- and pro-longevity genes. Our analysis of pathways common to aging-related genes allows us to systematically classify pathways as anti- or pro-longevity, even though these mostly recapitulate previous findings. By contrast, our comparison of aging-related genes with age-related disease genes reveals several unexpected results: we found an association between aging-related genes and age-related diseases, yet this association is surprisingly organism-specific and driven by a small cluster of genes. Besides, one major issue in network analysis is whether some genes being better studied than others (what we call publication bias) impacts the underlying datasets and subsequent results. We correct for publication bias and show that a small but detectable fraction of results from protein-protein interaction network analysis is indeed influenced by whether genes are more studied than others. Lastly, we identify and rank drugs being targeted by aging-related genes which are promising for additional studies.

Results

Our systematic analysis of the gerontome employed the GenAge database developed by our lab (6). This includes 298 human candidate aging-related genes and genes associated with aging and/or longevity in model organisms of which over 1,000 can be converted to human homologs (see Materials and Methods). Model organism aging-related genes were further classified as pro- or anti-longevity depending on their effects: pro-longevity genes are defined as genes whose decreased expression reduces lifespan and/or whose overexpression extends lifespan; accordingly, anti-longevity genes are those whose decreased expression extends lifespan and/or whose overexpression decreases it (6) (Materials and Methods). This work is the first to consider such classification in a systematic way.

Processes and pathways overrepresented in pro- and anti-longevity genes

First, we performed a functional enrichment analysis of pro- and anti-longevity genes in each of the major model organisms. For pro-longevity genes, the most significant enriched pathways were p53-signaling pathway and cell cycle in mice; hypoxia response via HIF activation in *Drosophila melanogaster*; regulation of autophagy and oxidative phosphorylation in *C. elegans* (Supplementary Dataset 1). On the other hand, for anti-longevity genes, insulin signaling, growth hormone signaling and IGF-1 receptor pathways were overrepresented in mice; the PI3 kinase pathway, oxidative phosphorylation and IGF pathway in *Drosophila*; oxidative phosphorylation, mTOR signaling pathway in *C. elegans*; ribosome in *Saccharomyces cerevisiae*. Some pathways like mTOR signalling, autophagy, insulin signalling and ribosome were enriched in more than one model organism (Supplementary Dataset 1).

In addition to the more traditional functional enrichment, we also used a recently proposed feature selection method, from the area of data mining (or machine learning) to select

relevant biological process Gene Ontology (GO) terms for predicting the pro-longevity or anti-longevity effect of a gene on a model organism (8). Among the top ranking GO terms identified by that feature selection method, terms associated with pro-longevity included apoptotic signalling pathway and cell cycle checkpoint in mice, lipid metabolic process in *Drosophila*, autophagy in *C. elegans* and telomere organization in *S. cerevisiae*. By contrast, top ranking GO terms associated with anti-longevity included positive regulation of multicellular organism growth in mice, sensory perception in *Drosophila*, and translation in *C. elegans* (Supplementary Dataset 2).

Although the two aforementioned methods work in very different ways, there is some overlap between their results. In particular, in the results for mice, both methods found terms related to insulin signaling or growth to be significantly associated with anti-longevity; and terms related to the cell cycle were found to be significantly associated with pro-longevity. In addition, some terms related to autophagy were found to be significantly associated with pro-longevity in *C. elegans* by both methods.

These results mostly recapitulate current knowledge of pathways associated with longevity manipulation in model organisms. Nonetheless, our results highlight pathways with pro- and anti-longevity effects and allow us to classify such pathways in a more consistent, systematic way.

Pro- and anti-longevity networks are intertwined

Next, we tested if aging-related genes interact with each other and if we can observe differences between the ways that pro- and anti-longevity genes form protein interaction networks. To perform this analysis, we employed protein-protein interaction data from BioGRID (see Materials and Methods) and focused our attention on worm genes, as the dataset

of aging-related genes in worms is by far the largest among the animal models in GenAge (Table 1).

Out of all the worm genes classified as anti- or pro-longevity genes ($n = 719$), 283 genes had interactions in the BioGRID interactome (109 pro- and 174 anti-longevity genes). The average connectivity degree of pro-longevity genes was slightly higher than that of anti-longevity genes (8.42 compared to 5.43), and on average, both sets of aging-associated genes included more connected genes than similarly-sized random sets from the interactome (the connectivity degree for the entire interactome is 3.8). The clustering coefficient of pro-longevity genes was also higher than that of anti-longevity genes (0.108 compared to 0.063), showing that on average pro-longevity genes tend to cluster better than anti-longevity genes.

In addition, we found that pro- and anti-longevity genes are much intertwined, with almost as many protein-protein interactions between genes from opposite categories (80 interactions) as between genes from the same category (43 interactions between pro-longevity genes and 56 interactions between anti-longevity genes) (Figure 1). While pro- and anti-longevity genes can form two network cores by themselves (28 genes are interconnected for each set), they also form a much larger network when taken together (90 genes), suggesting that the way in which pro- and anti-longevity genes determine lifespan is in many cases dependent on one another.

Various previous studies have shown aging-related genes to form strong networks (9, 10, 14), as is normal in biology, but our results expand these observations to pro- and anti-longevity subnetworks and support substantial interactions between pro- and anti-longevity genes.

Comparison with longevity-associated human genes

In addition to genetic manipulations in model organisms, a number of genes has been associated with longevity in human populations (20). We therefore also performed a functional enrichment analysis of these genes using data from the LongevityMap, which consists of 755 human genes, 328 of which associated with longevity in at least one genetic association study (20). For a first functional enrichment analysis using DAVID (see Materials and Methods), with background set as default, 49 clusters showed an enrichment score greater than 2.5 (Supplementary Dataset 3). With the background set as LongevityMap genes, 62 clusters with enrichment score > 2.5 were obtained (Supplementary Dataset 3). A similar functional annotation clustering pattern as the first run was observed (Supplementary Figure S1), and major enriched clusters consisted of: regulation of apoptosis, regulation of phosphorylation, response to environment, regulation of locomotion and response to hormone stimulus.

Results from human longevity-associated genes only modestly overlap with the above results for model organisms, although some pathways thought to be related to aging (e.g., apoptosis, response to oxidative stress and mTOR signaling) were found. Enriched clusters also included terms related to age-related diseases like cancer and diabetes mellitus (Supplementary Dataset 3). This may reflect how researchers choose candidate genes for longevity association studies, however. Perhaps researchers tend to select candidate genes for their studies that are suspected of playing important roles in human longevity, or in severe pathology processes that can significantly impair longevity.

Overlap between aging-related genes and age-related diseases

Next we aimed to study the genetic overlap between aging and age-related diseases (ARDs). For this analysis, we used human genes associated with ARDs from public databases (see Materials and Methods), human candidate aging-related genes and human homologs of

genes associated with aging in model organisms from GenAge (6). Common or shared genes between ARDs and aging gene sets are referred to as common aging and disease (CAD) genes. In addition to an analysis focused on individual age-related diseases, a set named "all diseases" and another named "all classes" were created, composed of all genes considered in the analyses per individual age-related disease and per age-related disease class, respectively (Materials and Methods).

As expected, the human aging-related gene set has the most associations with ARD genes. In addition, the immune system and respiratory tract disease classes only show a relation with aging in human aging-related genes (Figure 2A). Among the human homologs of genes associated with aging in model organisms, the musculoskeletal disease class only exhibits a significant overlap with aging-related genes in the mouse. A decrease in the overlap between aging and age-related diseases as evolutionary distance increases is also clear from our results (Figure 2A) with the mouse showing an overlap with more ARDs when compared with the other model organisms, even though fewer aging-related genes are known in mice than in flies or worms. This is also clear from looking at individual ARDs (Figure 3) since in mice there is a significant overlap with 9 and 17 ARDs for anti- and pro-longevity sets, respectively, while the remaining model organisms present the following values: *Drosophila* - 8 and 3; *C. elegans* - 5 and 0; and finally *S. cerevisiae* - 1 and 3.

We also compared overlaps between anti- and pro-longevity genes. Pro-longevity genes present a higher number of overlapping age-related diseases than anti-longevity genes for all orthologs (i.e., combining human orthologs of all genes from model organisms), mouse and *S. cerevisiae*. The opposite is verified for *Drosophila* and *C. elegans* sets (Figure 2A).

Supplementary Tables S6 and S7 include the p-values and the number of CAD-genes for, respectively, age-related disease classes and individual diseases. Neoplasms ($1.35\text{E-}56$), nutritional and metabolic ($9.70\text{E-}34$), cardiovascular ($2.00\text{E-}23$) and nervous system ($1.78\text{E-}18$)

classes have the strongest associations with human aging. There is an additional class not considered in individual ARD analysis, the eye diseases, which presents a positive association with aging only for human aging-related genes and mouse pro-longevity genes (Table S6).

Publication bias effects and correction

The inclusion of more and less studied genes may reduce the accuracy of the results. This is an issue when using large datasets that may contain systematic biases. Indeed, we observed a moderate correlation between the number of publications associated with a gene and its number of annotated protein-protein interactions (Spearman correlation coefficient = 0.67). While this is not unexpected, it could result in biases in systems biology analysis. To minimize this issue, a publication bias correction (PBC) based on the number of publications per gene was tested.

The first step of the PBC was setting a threshold for differentiation between more and less studied genes. Table 1 shows the average and the median number of publications computed for the following gene sets: human genome, human interactome, human aging-related genes and human homologs of aging-related genes from model organisms. The sets of aging-related genes have a higher average (range 13.0 to 30.3) and median (range 9 to 23) values when compared with the whole genome (average of 8.7 and median of 6) and the interactome (average of 10.4 and median of 8). This is expected but it shows that aging-related genes are more studied than average.

Thresholds between 8 and 20 publications were assessed in order to define the value with which the subsequent analyses were performed (see Supplementary Figure S2). Overall, we used 10 publications as a threshold.

Overlap between aging-related genes and age-related diseases with publication bias correction

The overlap analysis between aging-related genes and ARD genes was repeated after applying a filter for PBC (i.e., only containing genes with at least 10 publications). After PBC, the human aging-related gene set presents a significant overlap with ARD from all classes except for the immune system class (Figure 2B). For the human homologs of aging-related genes from model organisms, only the mouse and the *Drosophila* present significant overlaps with ARDs and the latter only presents significant results for the anti-longevity gene set. In the mouse, pro-longevity genes have a higher number of ARDs overlapping compared to anti-longevity genes. In the all orthologs gene set, both anti- and pro-longevity genes show an association with the neoplasms class. These results are supported by p-values in Supplementary Tables S8 and S9, also suggesting a general stronger overlap with aging of genes associated with neoplasms and nutritional and metabolic diseases.

Comparing the two analyses without and with PBC, respectively Figures 2A and 2B, we observe a decrease in the number of significant overlapping ARDs with all aging-related gene sets due to the exclusion of less studied genes. The human aging gene set is the least affected by the exclusion of less studied genes since after PBC it presents only a loss of 16% of its genes (298 to 253 genes). Small reductions are verified in small gene sets, such as the baker's yeast and the mouse. The opposite is verified in bigger gene sets, such as worms, which lose 47% and 50% of genes in anti- and pro-longevity sets, respectively. Finally, these same patterns are observed comparing Tables S6, S7, S8 and S9, which show statistical tests for the various overlaps.

Overlap between aging-related genes and age-related diseases in the interactome

The interactome (15000 genes) is a subset of the genome (20183 genes) within which only genes for which protein-protein interaction data is available are present (Table 1). We assessed the overlap between aging and ARDs genes when restricting the analysis to genes in the human interactome (Supplementary Figure S3). The distribution of aging-related gene overlaps with ARDs is similar in the interactome (Supplementary Figure S3) to the distribution in the whole genome (Figure 2A), both without and with PBC. The analyses present similarities in the total number of genes, the overlap significance (p-values), the number of CAD-genes, and the relations between age-related diseases and aging shown by the anti- and pro-longevity sets.

When comparing the whole genome and the interactome analyses without PBC (Tables S6, S7, S10 and S11) and the whole genome and the interactome analyses with PBC (Tables S8, S9, S12 and S13), there is a slight drop in the significance of overlaps with PBC, suggesting that some (but not all) results are indeed due to publication bias. Looking at the effect of PBC on the number of CAD-genes, there is again a slight decrease with PBC in the majority of ARD classes and individual diseases. We conclude that publication bias has a modest but noticeable effect on our results.

Since genes function in combination with other genes, studying genes and proteins as part of interaction networks is essential (21). To study the effect of genes which interact with aging-related genes in the links between aging and ARDs, we performed an analysis in which aging gene sets were composed of gene sets from the genome with PBC plus the genes directly connected with them (first order partners). This analysis revealed that aging-related gene sets including the first order partners are 18 to 51 times bigger than the original aging-related gene sets with PBC (Figure 4A). This increase in the number of genes is not proportional to the

initial gene set size, and human aging-related genes are the ones that interact more with other genes.

Regarding ARDs classes overlapping with aging-related genes, neoplasms and nervous system classes do so in all gene sets analysed. Cardiovascular, as well nutritional and metabolic classes are also present. Musculoskeletal diseases overlap with human aging-related genes and then they only overlap with anti-longevity genes of mice and *Drosophila*. Finally, respiratory tract diseases present a significant overlap with aging for the anti-longevity gene set in *C. elegans*.

Overall, there is a clear increase in the number of overlapping ARD genes with aging-related genes by including first order interaction partners, as well as in the number of CAD-genes, which is supported by a statistical significance analysis (Supplementary Tables S14 and S15). Given the large increase in the number of genes by including first order partners, these results are not surprising but they underscore the large interconnection of biological networks, including between aging and age-related diseases.

Co-expression network analysis

Co-expression networks offer a complementary perspective on biological interactions from protein-protein interaction networks. To study co-expression, data was downloaded from the GeneFriends database (22) and genes co-expressed with aging-related genes (human candidate genes plus human homologs of genes associated with aging in model organisms, all following PBC) were considered (see Materials and Methods). The inclusion of genes co-expressed with aging-related genes again changes the number of genes per set; and there is an increase which is proportional to the size of the initial set, i.e., larger sets have a greater increase in their sizes (Figure 4B).

At the level of age-related disease classes, neoplasms is the main class with a significant overlap between aging-related genes and ARDs genes, followed by the cardiovascular and nervous system classes. As before, the human aging-related gene set shows the highest association with ARDs. Immune system disease genes seem associated with aging-related genes when considering the anti-longevity genes of all orthologs together. Interestingly, there is a difference in overlapping classes between anti- and pro-longevity gene sets and, except in the mouse, anti-longevity genes present a higher relation to ARDs than pro-longevity genes, which is very clear in flies and worms (Figure 5).

In the mouse, the results show that pro-longevity genes slightly overlap with nervous system disease genes, but the association is verified due to just one disease. In *Drosophila*, anti-longevity genes are associated with nutritional and metabolic, cardiovascular and respiratory tract diseases, while the pro-longevity genes are associated only with nervous system diseases. Finally, there is no overlap with any age-related diseases in the pro-longevity gene set of worms, even though it contains more than two thousand genes. Moreover, there are a few CAD-genes (up to 15) which show non-significant overlaps with the assessed ARDs (Supplementary Tables S16 and S17).

Because including co-expressed genes increases the number of starting genes, there is an increase in the number of CAD-genes when including co-expressed genes. Using the human aging-related gene set as an example, there are 65 CAD-genes in the overlap with neoplasms genes but when including the co-expressed genes the overlap with neoplasms increases to 131 CAD-genes. However, the percentage of overlapping genes drops dramatically. For human aging-related genes, 22% and 0.06% are associated with ARDs, with and without co-expressed genes, respectively.

Properties of common genes between aging and age-related diseases

Common genes or CAD-genes from the overlap analyses can highlight clues about pathways which link aging to disease processes. CAD-genes were obtained from the overlap between the human aging set and ARDs genes from analyses per individual age-related diseases or per diseases class, both with PBC.

A small subset of aging-related genes are also associated with age-related diseases

The number of times that each CAD-gene overlaps with ARDs was defined as its frequency, and allows us to determine if some genes are involved in several diseases processes. Figure 6A shows the frequency of CAD-genes across all the age-related disease classes. A total of 94 genes were obtained from the overlap between the human aging-related genes and all the ARDs genes per class. A majority of these genes (approx. 83% genes) overlap with up to three classes (Figure 6A). Regarding genes which overlap with a great number of classes, *PON1* and *APOE* are at the top (Figure 6B), as well as some other genes involved in age-related changes, for example, *VEGFA*, *IL6* and *AR*. One gene present in all ARDs analyzed is *TNF* (tumour necrosis factor).

It is also interesting to explore aging-related genes which are not associated with any ARDs. The 94 CAD genes represent 37% of the human aging-related genes with PBC (253 genes), which means that most (63%) aging-related genes are not associated with any ARD class. From the perspective of ARDs genes (639 genes), about 15% have been related to human aging.

An analysis of the CAD-genes distribution was also performed by individual ARDs. A total of 90 genes were found to overlap between human aging-related genes and ARD genes. Figure 7A shows that the number of genes involved in several ARDs is small, and about 59% (53 genes) of the 90 genes are associated with up to three ARDs. The pattern of distribution is

similar to the analysis by ARD class and *TNF*, *PON1*, *APOE* and *VEGFA* are the top of CAD-genes among ARDs for both analyses (per age-related disease class and per individual disease) (Figure 7B). In this analysis, the percentage of aging-related genes not associated with any age-related disease is about 64%. Similar to above, from the perspective of ARD genes (596 genes), only 15% have been related to human aging.

Pathways and processes linking aging and age-related diseases

A functional enrichment analysis was performed in CAD-genes. The background used was the set of human aging-related genes plus ARDs sets. Functional enrichment for CAD-genes from all analyzed ARD classes shows that these genes are associated with: negative regulation of apoptosis, cell cycle, positive regulation of DNA, positive regulation of protein metabolic process and response to stimuli (Supplementary Table S1).

Looking at CAD-genes in individual ARDs, only neoplasms, nutritional and metabolic, musculoskeletal and eye classes have significant functional clusters. Relatively to neoplasms, CAD-genes are more associated with negative regulation of apoptosis, DNA repair, regulation of cell cycle and cancer, which is in line with cancer etiology and its relationship to aging (17). CAD-genes from the nutritional and metabolic class are related to response to insulin stimulus and positive regulation of lipid process, while CAD-genes for musculoskeletal diseases only show an association with the extracellular region. Finally, eye diseases CAD-genes seem to be associated with positive regulation of RNA metabolic process (Supplementary Table S1).

Increased network connectivity in genes common to aging and age-related diseases

Network approaches consider as a measure of node (gene) relevance the node's degree, which represents the number of connections of each node. This measure helps to define

hubs, which in general are deemed essential nodes with many connections. To understand if CAD-genes are likely to be hubs, a comparison between the degree of CAD-genes and ARD genes or aging-related genes (the non-common genes) was made using protein-protein interaction data (Supplementary Table S2).

Age-related disease class analysis shows significant differences in node degree between CAD-genes and controls (p -value < 0.001). The median node degree of CAD-genes (47) is substantially higher than the median for the control set (11). Looking at ARD classes, only two classes have a significant ($p < 0.05$) difference between the two sets investigated: neoplasms and immune system diseases. Neoplasms present a higher median for CAD-genes (47) compared to the control set (23.5), while for the immune system class the opposite is verified (8.5 vs 43) (Supplementary Table S2).

The results from the analysis per individual ARDs show a significant difference in the number of node connections for four diseases: atherosclerosis ($p = 0.002$), breast neoplasm ($p = 0.020$), hypersensitivity ($p = 0.019$) and osteoporosis ($p = 0.039$). Except for breast neoplasm, the median for CAD-genes is lower when compared to the ARD genes (Supplementary Table S2).

Processes associated with aging-related genes not associated with age-related diseases

A functional enrichment was performed for genes from the human aging gene set which are not associated with any ARD. The main processes in the functional enrichment are: response to DNA damage, negative regulation of apoptosis, ATP-binding, negative regulation of transcription, DNA repair, aging, telomere maintenance, response to several stimuli, negative regulation of gene expression, cancer and signalling pathways (for examples, insulin, IL3 and MAPKinase). Of these terms, the ones with the higher cluster scores are response to

DNA damage and negative regulation of apoptosis. The full list of significantly enriched terms is in Supplementary Table S3.

Molecular evolutionary rates of aging- and disease-related genes

Aging-related and disease-related genes are also known to differ from the genome-wide average at the level of selection pressures. The study this, the dN/dS ratio of the human aging-related genes and the ARD genes sets was analysed and compared to the remaining genome (see Materials and Methods). Results show a significant ($p < 0.001$) difference between ARD genes and the other genes in the genome, wherein ARD genes have a higher median dN/dS ratio (0.137) than the whole genome (0.091). Although there was a difference between aging (median of 0.079) and non-aging genes (median of 0.093), this was not statistically significant ($p\text{-value} = 0.155$).

The dN/dS ratio was also assessed in anti- and pro-longevity genes. A difference in dN/dS ratio between anti- and pro-longevity genes was only observed in *C. elegans* ($p\text{-value} = 0.046$, which is not significant after Bonferroni correction), so we find no evidence of differences in molecular evolution rates between anti- and pro-longevity genes.

Searching for patterns and features which could define CAD-genes, their molecular evolutionary (dN/dS) rate was analysed in comparison with aging-related genes and ARD genes. The CAD-genes used were from the overlaps of the three ARD classes with more genes: all classes together, neoplasms and nutritional and metabolic diseases. No statistically significant differences were found, suggesting that molecular evolution rates of CAD-genes are not different from other aging and ARD genes.

Drugs predicted from aging-related gene interactions with drugs

Given the large number of aging-related genes and pathways identified, there is great interest in identifying drugs that target them and may potentially have clinical benefits (23). To obtain candidate drugs affecting the aging process, we employed publicly available drug-gene interaction data (see Materials and Methods). In total, 376 drugs whose targets overlapped with aging-related genes were obtained. Twenty statistically-significant drugs that have more interactions with aging-related genes than expected by chance were obtained after Bonferroni correction (Supplementary Table S4).

The majority of the drugs obtained from this analysis were histone deacetylase inhibitors used for the treatment of cancer. This might be due to an overrepresentation of cancer drugs in public databases. Nonetheless, three known lifespan-extending drugs were identified: sodium phenylbutyrate, valproic acid and everolimus (Supplementary Table S4). The fact that experimentally validated aging-related drugs are detected by our methodology suggests that this approach may be useful to identify new candidate drugs with effects on aging.

Discussion

To our knowledge, ours is the largest analysis of the gerontome to date, and the first to consider pro- and anti-longevity genes in a systematic fashion. We first characterized functions and pathways overrepresented in pro- and anti-longevity genes. Major anti-longevity pathways and processes include insulin signaling, growth hormone signaling and mTOR signaling. Key pro-longevity pathways include p53, cell cycle and autophagy. Although such pathways and processes are known to be related to aging (2, 4, 5, 24), it is interesting that they are classified as anti-and pro-longevity in our systematic analysis of the genetics of aging. Differentiation between anti-longevity and pro-longevity genes and processes can provide additional clues about aging-related processes and can help identify other genes with a similar effect on aging.

In order to find relations between aging and ARD, we compared aging-related gene sets with ARD genes. Limitations of our study include the fact that possibly many genes associated with longevity and disease remain to be identified, and the causal genes in many genetic associations with disease are still unknown. In spite of these caveats, our results show an association between aging and ARDs at the genetic level, although this is surprisingly species-specific with a stronger overlap in mice than in invertebrates (flies and worms) and practically no overlap in yeast.

The overlap analyses of anti- and pro-longevity genes shows differences in musculoskeletal, nervous system and cardiovascular diseases. The identified overlaps suggest that the musculoskeletal and nervous systems are related to pro-longevity genes while anti-longevity genes seem more associated with cardiovascular diseases. Looking at ARD classes which overlap with human aging-related genes, a significant overlap is verified for all classes as expected, except for immune system diseases in the analysis with PBC. The nutritional and metabolic diseases, the neoplasms, the cardiovascular diseases and the nervous system diseases

have the most significant overlap with human aging-related genes. Eye diseases, respiratory tract diseases (which we considered a negative control) and immune system diseases had the least overlap, but it is important to mention that these are (together with musculoskeletal diseases) the age-related disease classes with fewer genes (Figure 8).

Genes historically associated with diseases are more likely to be studied. A publication bias correction approach, based on the number of publications associated with each gene, was applied in order to explore and reduce such biases. The analyses with and without PBC, when compared, show the effect of the removal of less studied genes (Figures 2A vs. 2B). The overlaps for *C. elegans* and *S. cerevisiae* disappear when the PBC is applied, which supports the hypothesis that some overlaps are statistically significant only due to an overrepresentation of better-studied genes. The comparison of analyses with and without PBC proves that systematic researcher biases can influence the results in large-scale systems biology, genomic and genetic analysis.

From our network analysis including the first order protein-protein interaction partners, it is possible to conclude that aging-related genes are widely connected to other genes, which is supported by the huge increase in gene sets' sizes (Figure 4A). There is an increase in the number of CAD-genes (the common or shared genes by aging and ARD) when including the first order partners, which suggests widespread interactions between aging-related genes and genes associated with age-related diseases. The results are also in agreement with recent research using genome-wide association studies (GWAS) data, which showed the same conclusion for five age-related categories: neurodegenerative, cancers, cardiovascular, metabolic and other diseases (25). A co-expression analysis of the links between aging and ARDs supports the idea of species-specific effects, but with more anti-longevity genes in invertebrates being related to ARDs. It is tempting to speculate that perhaps anti-longevity genes work together more tightly in transcriptional networks than pro-longevity genes.

Previous studies of the association between aging and diseases have demonstrated that the association is established by a small number of genes (25). Indeed, in the present analysis, CAD-genes represent a minority of the aging-related genes. CAD-genes are mainly related to apoptosis, metabolic regulation and DNA damage. These processes are similar to those previously reported to be associated with aging and may hint at underlying mechanisms important in various age-related diseases. CAD-genes also showed a higher number of connections with other genes than the remaining genome, which suggests that those genes tend to be hubs in networks. *TNF*, *PON1*, *APOE* and *VEGFA* are present in a great number of ARDs, which is in line with their involvement in some of the essential pathways whose disruption compromises metabolism and can lead to pathologies (26, 27).

The dN/dS ratio analysis showed a statistically significant higher dN/dS ratio of ARD genes when compared to the remaining genome, while aging-related genes had a lower dN/dS ratio that was not statistically significant. Therefore, we can affirm that ARD genes have a higher predisposition to changes in their sequence than aging-related genes. These results are in line with previous findings: an analysis using a previous version of GenAge found that aging-related genes have a lower dN/dS ratio (28). One previous study found a higher molecular evolutionary rate in disease genes (29). Our results further suggest that aging-related genes tend to be evolutionarily conserved, perhaps because they are part of essential pathways and conserved pleiotropic effects on aging (28), while genes associated with age-related diseases may be under relaxed selection given that they impact later in life.

Finally, taking advantage of a database of gene-drug interactions, we mapped GenAge's genes to drugs and obtained a list of 19 candidate drugs for aging effects. Of these, three are already experimentally validated and the rest is yet to be explored. As such, these compounds are promising for future studies.

Concluding remarks

The main conclusion from this work is that aging and age-related diseases are related and share more genes than expected by chance. Human aging-related genes showed a considerable overlap with ARDs. These overlaps are driven by a small subset of aging-related genes which are associated with various age-related diseases and are hubs in networks. Besides, the extent of overlaps decreases with the increase of evolutionary distance, and yeast aging-related genes show practically no overlap with ARDs. Novel differences in overlapping age-related disease classes between anti- and pro-longevity genes were observed: Nervous system and musculoskeletal diseases seem more associated with pro-longevity, while cardiovascular diseases have a stronger association with anti-longevity genes. Moreover, network analyses (protein-protein interactions (PPI) and co-expression) suggest the existence of intermediate genes which promote the associations between aging and age-related disease genes. Overall, our work establishes a new standard in the analysis of aging-related genes in a systematic way.

Materials and Methods

Aging- and Longevity-Associated Genes

Aging-associated genes were obtained from GenAge Build 17 (6). These include 298 human candidate aging-related genes. GenAge also includes aging- or longevity-related genes in model organisms. For use with human datasets, human orthologs of model organisms genes were used, composed of 1037 genes from the four main biomedical model organisms: mouse, fruit fly, roundworm and baker's yeast. The genes of each model organism were separated by their longevity classification: anti- or pro-longevity. Pro-longevity genes are genes whose decreased expression (due to knockout, mutations or RNA interference) reduces lifespan and/or whose overexpression extends lifespan; conversely, anti-longevity genes are those whose decreased expression extends lifespan and/or whose overexpression decreases it (6). Genes which were not included in one of these two longevity classes were excluded. A small number (19) of genes with both anti- and pro-longevity classifications were also excluded.

To sum up the data, aging-related genes were divided into 11 gene sets: one set of 298 human aging-related genes, two sets (anti- and pro-longevity) of human orthologs from each model organism and two sets with all human orthologs of genes in all model organisms. The mouse sets (i.e. human orthologs of genes associated with aging in mice), have 23 and 59 genes, the fruit fly sets have 48 and 87 genes, the roundworm sets have 381 and 290 genes, and lastly the baker's yeast sets have 41 and 13 genes, respectively anti- and pro-longevity (Supplementary Table S18). Finally, sets with all orthologs have 448 and 421 genes, anti- and pro-longevity sets, respectively. The full lists of human aging-related genes and human orthologs are available in the supplementary material (Supplementary Dataset 4).

Data on human genes associated with longevity in genetic association studies were obtained from the LongevityMap build 1 (20). In the full set of 755 genes, there were 328 genes with at least one significant result reported.

Age-Related Diseases Genes

Age-related disease (ARDs) genes were assembled on 15-04-2015 from a diseases list compiled by a National Institute of Aging study. The list only includes the genes with a positive association with the disease phenotype and with a MeSH annotation (30). This list is available online (https://www.irp.nia.nih.gov/branches/rrb/dna/gene_sets.htm) and it was compiled using information from the Genetic Association Database (30).

The original list includes many diseases not relevant for the present analysis since our interest focuses on complex ARDs. To select relevant ARDs, diseases with fewer than 20 genes associated and diseases of non-age-related diseases classes were excluded. We chose a threshold of 20 genes because it captures the major age-related diseases yet not so many diseases that our findings end up being diluted (Table S19). The original list also includes processes and conditions, for example, insulin resistance and hyperlipidemia, which are dysfunctions, and for that reason were also excluded. The following analysed classes were described as age-related in the literature: cardiovascular diseases, eye diseases, immune system diseases, musculoskeletal diseases, nervous system diseases, nutritional and metabolic diseases and neoplasms (2, 31). Respiratory tract diseases were considered as negative controls since the two diseases (after application of the described selection criteria) in this class are asthma, which is not considered an age-related disease (32), and chronic obstruction pulmonary disease, which is primarily environmental.

Selection of individually studied ARDs was made based on two criteria: first, the number of genes, to have larger sample sizes and increase statistical power. The second

criterion was how often and common in each disease. An example of selection is the case of ovarian neoplasm, which presents a smaller number of genes and is better known than head and neck neoplasms. In order to have a representative selection, seven diseases classes were included in the individual age-related disease analysis; for classes with a large number of diseases, we selected the top five or six most representative individual diseases of the class. Eye diseases were excluded from the individual disease analysis, since they include only non-common diseases with a small number of genes. Diseases that are primarily driven by environmental factors, like chronic obstruction pulmonary disease, were also not studied.

In total, 893 different genes associated with ARDs were considered. Figure 8 shows the number of genes per ARD class and the number of genes shared with each one of the remaining classes. The list of ARDs used in the present analysis is summarized in Supplementary Table S20. In total, 40 diseases were part of ARD classes, of which 22 ARDs were analysed individually. The list of all age-related diseases and their related genes is available in the supplementary material (Supplementary Dataset 5). Our full datasets are also available on GitHub (<https://github.com/maglab/genage-analysis>).

Protein-Protein Interaction and Gene Co-expression Data

Protein-protein interactions were obtained using the BioGRID plug-in available in Cytoscape, on 16-04-2015, by downloading the available node and edge tables. The two main types of interactions ("physical association" and "direct interaction") represent 124,238 of 140,891 interactions, involving 14,721 of 15,000 proteins. As such, the interactome analysis was performed using the full interactome. To obtain the first-order partners of aging-related genes, a Python script was used to compile connections between all genes in the interactome and return the merged list of seed genes and genes connected to them.

Co-expression data from RNA-Seq was obtained using GeneFriends (22) on 03-06-2015. To obtain the co-expressed genes, a significance threshold of $2.5E-06$ was applied to the p-value retrieved from GeneFriends. The threshold was defined by correction of standard α (0.05) using a Bonferroni correction where N represents the genome size (20183 genes).

Publication Bias Correction (PBC)

The number of publications was compiled from the Swiss-Prot PubMed annotation list, downloaded on 23-04-2015. Only human and reviewed genes were considered for this analysis. Although PubMed publications annotated in Swiss-Prot are not the total number of publications for each entry, they represent a curated selection. Thus, Swiss-Prot was selected as the source for the number of publications due to its curated nature, which makes it a reliable source of annotated data for protein coding genes.

Gene Features

The list of all human genes was collected from GenBank on 31-01-2015. For this analysis, only human annotated and protein coding genes were considered, which represent a set of 20,183 genes.

Gene length was computed using the start and end position in the genome from GenBank, downloaded on 15-05-2015. Protein length and mass were obtained from Swiss-Prot. Only reviewed genes were considered.

Molecular evolutionary rate (dN/dS) was calculated from the number of synonymous (dN) and non-synonymous (dS) substitutions downloaded from Ensembl BioMart, selecting the Ensembl Genes 80 database and *Homo sapiens* genes (GRCh38.p2) dataset on 17-05-2015. Of all the model organisms considered in the present work, only mouse orthologs present dN and

dS values, due to the great evolutionary distance shown by the other organisms. Thus, all dN/dS ratio comparisons consider the evolution between mice and humans.

Feature Selection Method

We used a recently proposed feature selection method, from the area of data mining (or machine learning) to select the relevant Gene Ontology (GO) terms for predicting the pro-longevity or anti-longevity effect of a gene on a model organism (8). In essence, we addressed the classification task of data mining, where the goal is to predict the class (pro- or anti-longevity effect) of an instance (a gene) based on predictive features (GO terms) associated with that gene. The used feature selection method differs from other feature selection methods typically used in data mining in two important ways, as follows. First, it selects a specific set of features relevant for the classification of each instance, instead of selecting the same set of features for all instances, as usual in data mining. This increases the flexibility of the feature selection process, recognizing that the optimal set of GO terms for predicting the pro- or anti-longevity effect of a gene varies across different genes. Second, the used method performs “hierarchical” feature selection, in the sense that it takes into account the hierarchical structure of the GO in order to improve the feature selection process; unlike conventional (“flat”) feature selection methods.

That feature selection method was applied to datasets with data about aging-related genes from the four traditional model organisms, namely: mouse, fly, roundworm and yeast. The results of the feature selection method were transformed into a rank of GO terms as follows. For each dataset (model organism), for each GO term, we counted the number of instances (genes) where that GO term was one of the relevant features selected by the feature selection method for predicting the pro- or anti-longevity effect of the gene. Then, we ranked the GO terms in decreasing order of this frequency of selection. We also used a statistical test

of significance based on the binomial distribution to detect which GO terms were significantly associated with the class being predicted. A detailed description of the feature selection method and how its results were used to rank the GO terms can be found in (8).

Overlap Analysis

A significant overlap between aging-related genes and ARDs is defined as: i) an observed number of CAD-genes above the number of CAD-genes expected by chance; and ii) a p-value below 0.05 (Fisher's exact test). Genome and interactome analysis used whole genome and interactome as background. The background for the first order partners analysis was the seed list plus the interactome and was adjusted for each aging gene set since the seed list varies between different aging sets. In this analysis, the interactome was added to the background since the first order partners were selected from that group.

Functional Enrichment Analysis

Functional enrichment analysis using the Database of Annotation Visualization and Integrated Discovery (DAVID) (33) was performed to identify overrepresented categories. The analysis was done by running the Functional Annotation Clustering module under default parameters. Unless otherwise stated, the whole genome was used as background. Enrichment scores (E. Score) above 1.3 (which corresponds to $p = 0.05$) are widely accepted as relevant (33); however, in this analysis a threshold of 2.5 (corresponding to $p = 0.003$) was used for more significant results. A Benjamini correction was applied for correcting for multiple hypothesis testing.

Candidate Drugs from GenAge Targets

To identify candidate drugs with possible anti-aging properties, the Drug Gene Interaction Database (DGIdb) version 2 (34) was used. We classified all 44 types of drug-gene interactions in DGIDB into either “Anti” (decrease gene expression or activity) or “Pro” (increase gene expression or activity) or “Neither” (non-applicable or undefined effects), so that they can be matched with GenAge genes to obtain a putative lifespan-extending effect (Supplementary Table S5). Drugs were obtained by considering if they interact with GenAge genes in a way that would be predicted to extend lifespan. That is, for an “Anti” drug, only the interactions with anti-longevity genes are scored; and vice-versa for “Pro” drugs.

In total, 376 drugs were obtained which were ordered based on ascending p-value obtained using a one-tailed hypergeometric test. After Bonferroni correction ($\alpha=0.05$), 20 statistically significant drugs were obtained.

Acknowledgements

We thank Louise Crompton for assistance in compiling the relevant literature. GenAge is supported by a Wellcome Trust grant (104978/Z/14/Z) to JPM.

Conflict of interest statement

None declared.

References

- 1 Franceschi, C., Bonafe, M., Valensin, S., Olivieri, F., De Luca, M., Ottaviani, E. and De Benedictis, G. (2000) Inflamm-aging - An evolutionary perspective on immunosenescence. *Molecular and Cellular Gerontology*, **908**, 244-254.
- 2 Lopez-Otin, C., Blasco, M.A., Partridge, L., Serrano, M. and Kroemer, G. (2013) The Hallmarks of Aging. *Cell*, **153**, 1194-1217.
- 3 de Magalhaes, J.P. (2011) Stuart-Hamilton, I. (ed.), In *An Introduction to Gerontology*. Cambridge University Press, Cambridge, pp. 21-47.
- 4 de Magalhaes, J.P. (2005) Open-minded scepticism: inferring the causal mechanisms of human ageing from genetic perturbations. *Ageing Res Rev*, **4**, 1-22.
- 5 Kenyon, C.J. (2010) The genetics of ageing. *Nature*, **464**, 504-512.
- 6 Tacutu, R., Craig, T., Budovsky, A., Wuttke, D., Lehmann, G., Taranukha, D., Costa, J., Fraifeld, V.E. and de Magalhaes, J.P. (2013) Human Ageing Genomic Resources: Integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res*, **41**, D1027-1033.
- 7 Kwon, J., Lee, B. and Chung, H. (2010) Gerontome: a web-based database server for aging-related genes and analysis pipelines. *BMC Genomics*, **11 Suppl 4**, S20.
- 8 Wan, C., Freitas, A.A. and de Magalhaes, J.P. (2015) Predicting the Pro-Longevity or Anti-Longevity Effect of Model Organism Genes with New Hierarchical Feature Selection Methods. *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, **12**, 262-275.
- 9 de Magalhaes, J.P. and Toussaint, O. (2004) GenAge: a genomic and proteomic network map of human ageing. *FEBS Lett*, **571**, 243-247.
- 10 Zhang, Q., Nogales-Cadenas, R., Lin, J.R., Zhang, W., Cai, Y., Vijg, J. and Zhang, Z.D. (2016) Systems-level analysis of human aging genes shed new light on mechanisms of aging. *Hum Mol Genet*.
- 11 Freitas, A.A., Vasieva, O. and de Magalhaes, J.P. (2011) A data mining approach for classifying DNA repair genes into ageing-related or non-ageing-related. *BMC Genomics*, **12**, 27.

- 12 Faisal, F.E. and Milenkovic, T. (2014) Dynamic networks reveal key players in aging. *Bioinformatics*, **30**, 1721-1729.
- 13 Wuttke, D., Connor, R., Vora, C., Craig, T., Li, Y., Wood, S., Vasieva, O., Shmookler Reis, R., Tang, F. and de Magalhaes, J.P. (2012) Dissecting the gene network of dietary restriction to identify evolutionarily conserved pathways and new functional genes. *PLoS Genet*, **8**, e1002834.
- 14 Tacutu, R., Shore, D.E., Budovsky, A., de Magalhaes, J.P., Ruvkun, G., Fraifeld, V.E. and Curran, S.P. (2012) Prediction of *C. elegans* longevity genes by human and worm longevity networks. *PLoS One*, **7**, e48282.
- 15 van Dam, S., Cordeiro, R., Craig, T., van Dam, J., Wood, S.H. and de Magalhaes, J.P. (2012) GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics*, **13**, 535.
- 16 Avramopoulos, D., Szymanski, M., Wang, R.H. and Bassett, S. (2011) Gene expression reveals overlap between normal aging and Alzheimer's disease genes. *Neurobiology of Aging*, **32**.
- 17 de Magalhaes, J.P. (2013) How ageing processes influence cancer. *Nat Rev Cancer*, **13**, 357-365.
- 18 Peysselon, F. and Ricard-Blum, S. (2011) Understanding the biology of aging with interaction networks. *Maturitas*, **69**, 126-130.
- 19 Wang, J.G., Zhang, S.H., Wang, Y., Chen, L.N. and Zhang, X.S. (2009) Disease-Aging Network Reveals Significant Roles of Aging Genes in Connecting Genetic Diseases. *Plos Computational Biology*, **5**.
- 20 Budovsky, A., Craig, T., Wang, J., Tacutu, R., Csordas, A., Lourenco, J., Fraifeld, V.E. and de Magalhaes, J.P. (2013) LongevityMap: a database of human genetic variants associated with longevity. *Trends Genet*, **29**, 559-560.
- 21 de Magalhães, J. and Tacutu, R. (2016) Kaeberlein, M. and Martin, G. (eds.), In *Handbook of the Biology of Aging, 8th edition*. Academic Press, London, UK, pp. 263-285.
- 22 van Dam, S., Craig, T. and de Magalhaes, J.P. (2015) GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic Acids Res*, **43**, D1124-1132.

- 23 de Magalhaes, J.P., Wuttke, D., Wood, S.H., Plank, M. and Vora, C. (2012) Genome-environment interactions that modulate aging: powerful targets for drug discovery. *Pharmacol Rev*, **64**, 88-101.
- 24 de Magalhaes, J.P. and Faragher, R.G. (2008) Cell divisions and mammalian aging: integrative biology insights from genes that regulate longevity. *Bioessays*, **30**, 567-578.
- 25 Johnson, S.C., Dong, X., Vijg, J. and Suh, Y. (2015) Genetic evidence for common pathways in human age-related diseases. *Aging Cell*, **14**, 809-817.
- 26 Kumar, D. and Rizvi, S.I. (2014) Age-Dependent Paraoxonase 1 (PON1) Activity and LDL Oxidation in Wistar Rats during Their Entire Lifespan. *Scientific World Journal*.
- 27 Alber, H.F., Frick, M., Dulak, J., Dorler, J., Zwick, R.H., Dichtl, W., Pachinger, O. and Weidinger, F. (2005) Vascular endothelial growth factor (VEGF) plasma concentrations in coronary artery disease. *Heart*, **91**, 365-366.
- 28 de Magalhaes, J.P. and Church, G.M. (2007) Analyses of human-chimpanzee orthologous gene pairs to explore evolutionary hypotheses of aging. *Mech Ageing Dev*, **128**, 355-364.
- 29 Smith, N.G.C. and Eyre-Walker, A. (2003) Human disease genes: patterns and predictions. *Gene*, **318**, 169-175.
- 30 Zhang, Y., De, S., Garner, J.R., Smith, K., Wang, S.A. and Becker, K.G. (2010) Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. *BMC Med Genomics*, **3**, 1.
- 31 Loeser, R.F. (2010) Age-Related Changes in the Musculoskeletal System and the Development of Osteoarthritis. *Clinics in Geriatric Medicine*, **26**, 371-+.
- 32 de Marco, R., Locatelli, F., Sunyer, J., Burney, P. and Surve, E.C.R.H. (2000) Differences in incidence of reported asthma related to age in men and women - A retrospective analysis of the data of the European Respiratory Health Survey. *American Journal of Respiratory and Critical Care Medicine*, **162**, 68-74.
- 33 Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4**, 44-57.

34 Griffith, M., Griffith, O.L., Coffman, A.C., Weible, J.V., McMichael, J.F., Spies, N.C., Koval, J., Das, I., Callaway, M.B., Eldred, J.M. *et al.* (2013) DGIdb: mining the druggable genome. *Nature Methods*, **10**, 1209-1210.

Legends to figures

Figure 1: Protein-protein interactions between aging-related genes. Pro-longevity genes are depicted in red and anti-longevity genes in green. For each of the two gene sets, the smaller inside ellipse indicates genes that form a continuously connected network. Left right straight and curved arrows are used to summarize undirected interactions between genes from different and the same gene set, respectively.

Figure 2: The y-axis quantifies the number of age-related diseases which significantly overlap with aging-related genes; the x-axis describes the aging-related gene sets studied according to the source organism (i.e., human homologs of aging-related genes from each model organism). The columns have seven different colours to represent each age-related disease classes analysed: Neoplasms (light blue), Nutritional and Metabolic diseases (orange), Nervous System diseases (light grey), Cardiovascular diseases (yellow), Musculoskeletal diseases (blue), Respiratory Tract diseases (green) and Immune System diseases (dark blue). The first column represents the number of age-related diseases with a significant overlap with candidate human aging-associated genes. Model organisms are ordered by evolutionary proximity to humans. The genome was considered as background. The secondary x-axis displays the number of genes from the respective gene sets. **Figure 2a** shows the number of significant overlapping aging-related genes with age-related diseases. **Figure 2b** shows the number of significant overlapping aging-related genes with age-related diseases with PBC (i.e., only genes with more than 10 publications were used).

Figure 3: Overlapping aging-related genes for various organisms with age-related disease genes sets. Green means significant overlap between aging-related and age-related disease

genes and red means there is no significant overlap. Model organisms are in descending order of their proximity to humans.

Figure 4: The y-axis quantifies the number of age-related diseases which significantly overlap with aging-related genes; the x-axis describes the aging-related gene sets studied according to the source organism (i.e., human homologs of aging-related genes from each model organism). The columns have seven different colours to represent each age-related disease classes analysed: Neoplasms (light blue), Nutritional and Metabolic diseases (orange), Nervous System diseases (light grey), Cardiovascular diseases (yellow), Musculoskeletal diseases (blue), Respiratory Tract diseases (green) and Immune System diseases (dark blue). The first column represents the number of age-related diseases with a significant overlap with candidate human aging-associated genes. Model organisms are ordered by evolutionary proximity to humans. This analysis was performed with PBC. The secondary x-axis displays the number of genes from the respective gene sets. **Figure 4a** shows the number of significant overlapping aging-related genes with age-related diseases, including first order interaction partners. The interactome plus aging-related and age-related disease genes was considered as background. **Figure 4b** shows the number of significant overlapping aging-related genes with age-related diseases, including co-expressed genes. The genome was considered as background.

Figure 5: Overlapping aging-related genes and their co-expressed partners with age-related diseases for various classes for organisms. Green means there is at least one age-related disease from that class that significantly overlaps with aging-related genes and red means no association. Model organisms are in descending order of their proximity to humans. This analysis was performed without PBC.

Figure 6a CAD-genes distribution as associated with age-related disease classes. **Figure 6b** shows the genes involved in more than half of the seven possible disease classes. *TNF* is associated with all the age-related disease classes analyzed. This analysis was performed with PBC.

Figure 7a: CAD-gene distribution as associated with individual age-related diseases. **Figure 7b** shows the CAD-genes involved in more than ten individual diseases, with *PONI*, *TNF*, *APOE* the top 3 genes associated with the greatest number of age-related diseases. This analysis was performed with PBC.

Figure 8: Number of genes by age-related disease class (Total column) and shared with each other disease classes. The white cells present the number of genes shared between disease classes and the darker grey cells show the number of genes not shared with any other disease class.

Tables

Table 1 - Number of genes plus average and median number of publications per gene in each dataset

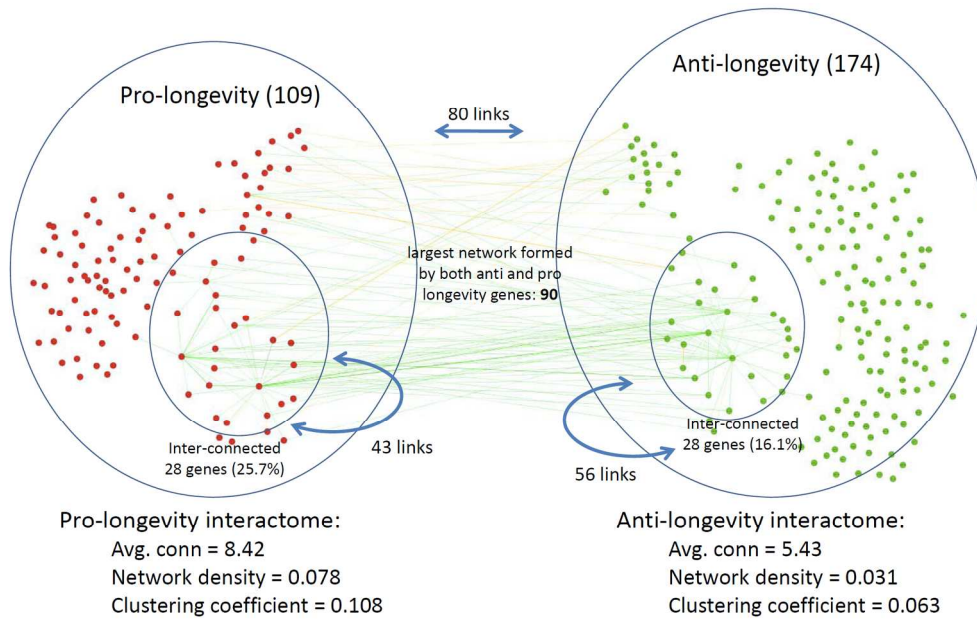
Dataset ^a	Num. of genes	Average num. pubs.	Median num. pubs.
Human genome (NCBI)	20183 ^b	8.7	6
Human interactome (BioGRID)	15000 ^c	10.4	8
Human aging-related genes	298	30.3	23
All aging-related orthologs	894	14.5	10
anti-longevity	448	13.2	9
pro-longevity	421	15.9	11
<i>M. musculus</i>	84	26.8	19
anti-longevity	23	22.7	13
pro-longevity	59	28.6	21
<i>D. melanogaster</i>	135	19.9	13
anti-longevity	48	20.1	12
pro-longevity	87	19.6	13
<i>C. elegans</i>	693	13.1	9
anti-longevity	381	13.0	9
pro-longevity	290	13.1	9
<i>S. cerevisiae</i>	62	17.9	14
anti-longevity	41	14.7	10
pro-longevity	13	23.2	15

Notes:

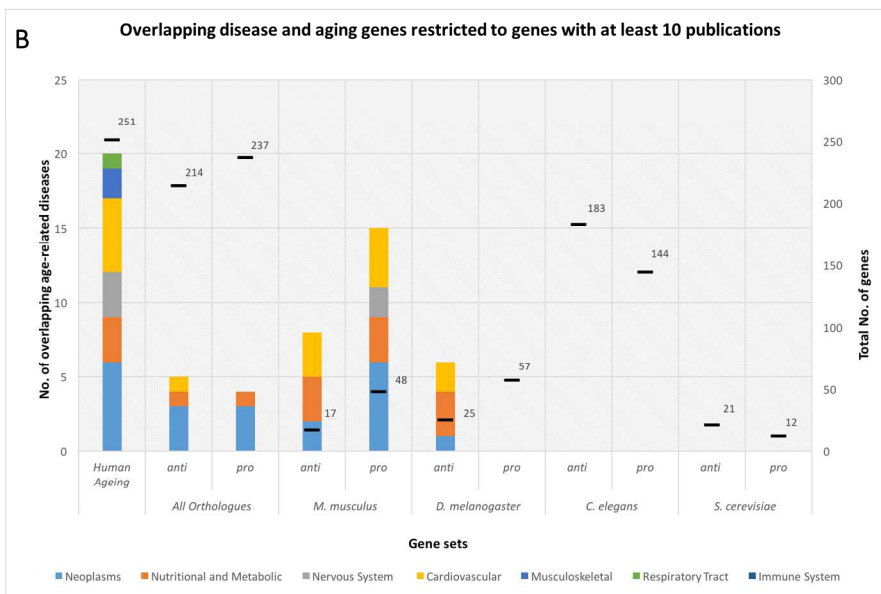
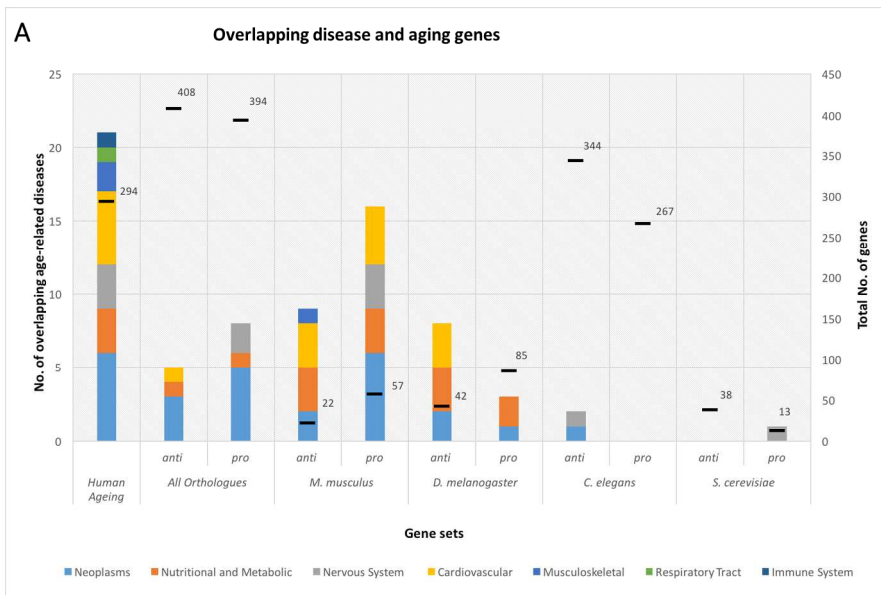
^a - All datasets refer to human genes, including human orthologs of genes from various model organisms

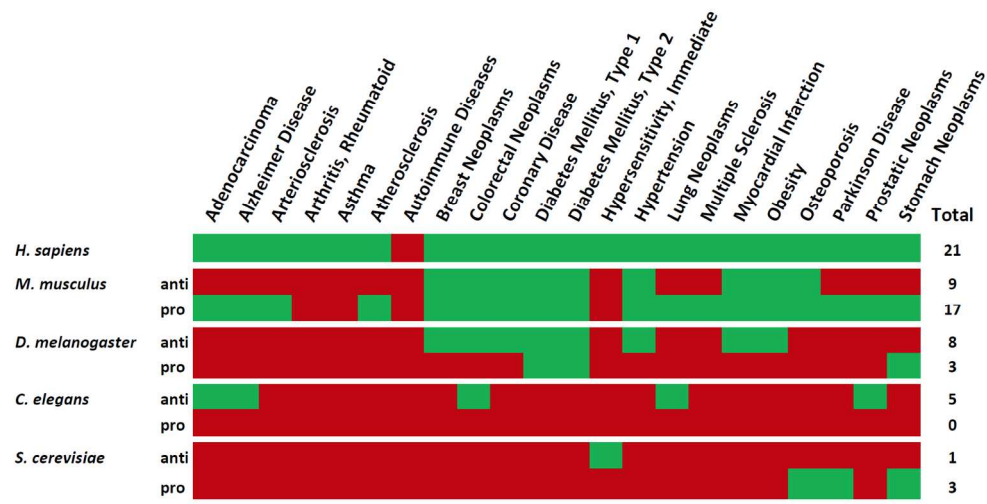
^b - Genome has 20183 annotated genes in NCBI but only 19071 are in the Swiss-Prot database

^c - Interactome has 15000 annotated genes in NCBI but only 14498 are in the Swiss-Prot database

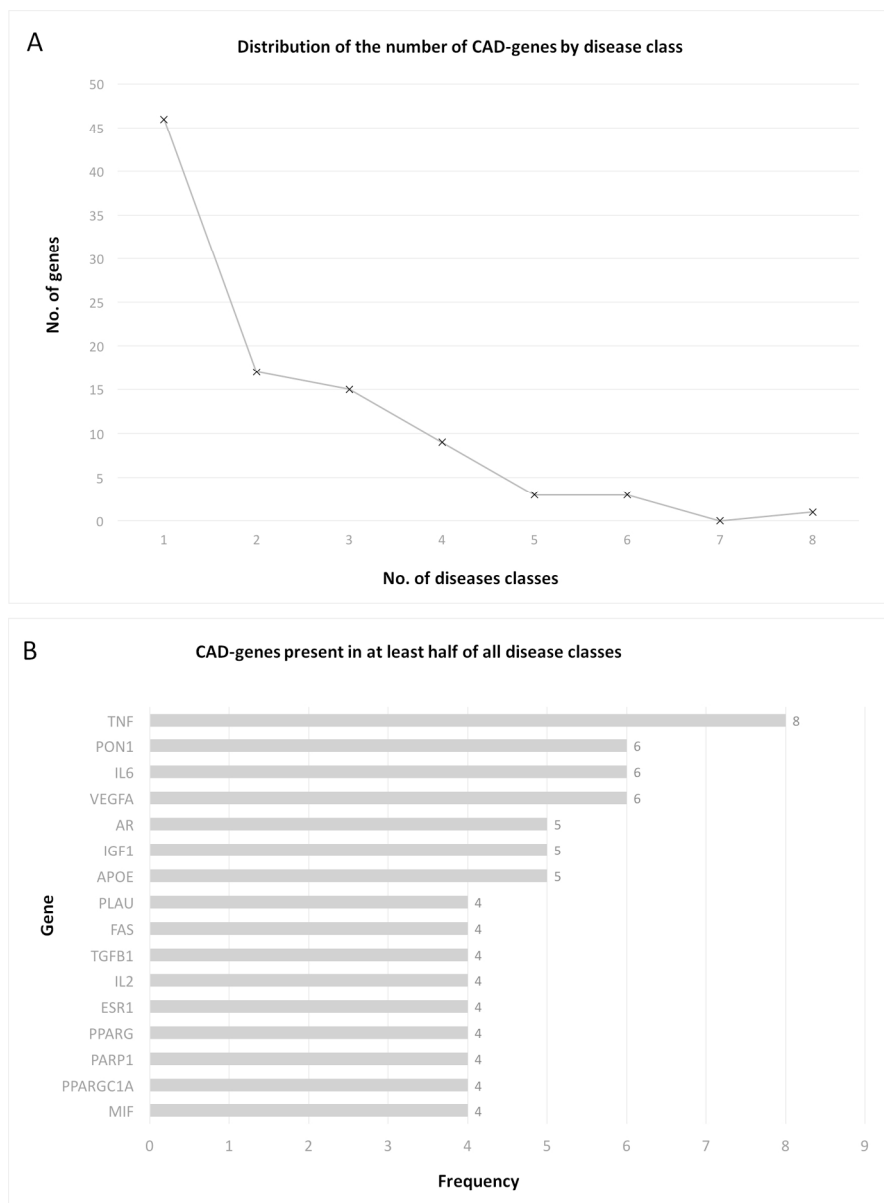


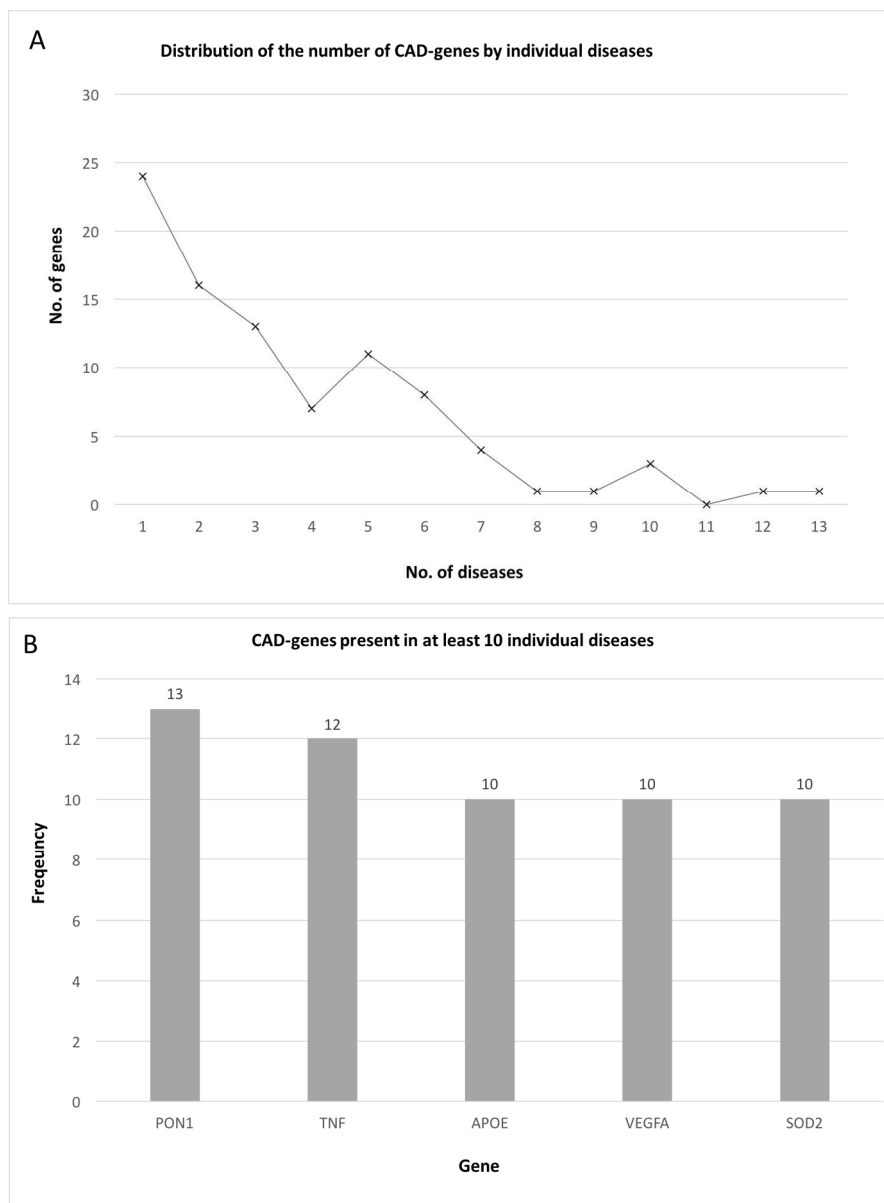
745x465mm (72 x 72 DPI)





		Cardiovascular	Immune System	Musculoskeletal	Neoplasms	Nervous System	Nutritional and Metabolic	Respiratory Tract
Human ageing		5	0	2	6	3	3	1
	Anti	4	1	0	6	2	3	0
All orthologues	Pro	0	0	0	6	0	0	0
	Anti	3	0	1	3	0	3	0
<i>M. musculus</i>	Pro	4	0	1	6	1	2	0
	Anti	2	0	0	0	0	3	1
<i>D. melanogaster</i>	Pro	0	0	0	0	2	0	0
	Anti	4	0	0	6	2	2	0
<i>C. elegans</i>	Pro	0	0	0	0	0	0	0
	Anti	0	0	0	0	0	0	0
<i>S. Cerevisiae</i>	Pro	0	0	0	0	0	0	0





	Cardiovascular	Eye	Immune System	Musculoskeletal	Neoplasms	Nervous System	Nutritional and Metabolic	Respiratory Tract	Total
Cardiovascular	115	37	17	49	84	60	140	45	320
Eye		15	17	26	31	27	50	25	75
Immune System			6	19	17	19	23	34	51
Musculoskeletal				23	45	43	46	30	107
Neoplasms					176	48	77	45	307
Nervous System						85	60	31	180
Nutritional and Metabolic							115	39	310
Respiratory Tract								49	144

550x415mm (72 x 72 DPI)